Joe Kelly
15209904 - Tutorial 2

Question 1

B.

Removed - “our, and, a, the, have, this, those, his, us, by”
Remained - "May”, "children", "children's", "thousand", "generations", "continue", "enjoy”, "benefits", "conferred", "upon", "united”, "country”, "cause", "yet", "rejoice”, "under", "glorious", "institutions", "bequeathed", "Washington", "compeers."

When using the wordcloud() function, it looks as if the the stop words have been removed as well as special characters by default. However, there has been no other manipulation to the data, such as stemming. The original sentence contains 42 words as shown in figure 1 code.

```
> sapply(gregexpr("\\W+", wordSentence), length) + 1
[1] 42
```

Figure 1: Counting all the words using R, the length is 42 words.

When the string is split into a character vector and unique number of words are counted, it contains 31, shown in Figure 2.

```
> library(stringr)
> stringWordsSplit <- str_split(wordSentence, pattern="\\s+")
> stringWordsSplit <- unlist(stringWordsSplit)
> unique(stringWordsSplit)
 [1] "May"          "our"          "children"     "and"          "children's"
 [6] "to"           "a"            "thousand"     "generations," "continue"
[11] "enjoy"        "the"          "benefits"     "conferred"    "upon"
[16] "us"           "by"           "united"       "country,"     "have"
[21] "cause"        "yet"          "rejoice"      "under"        "those"
[26] "glorious"     "institutions" "bequeathed"   "Washington"   "his"
[31] "compeers."
```

Figure 2: Number of unique words

This number is significantly less than the number of words that have appeared in the wordcloud. The word “to” appears 3 times in the original paragraph, but doesn’t appear in the wordcloud image shown below in figure 3. Other words that have been omitted from the wordcloud are “our, and, a, the, have, this, those, his, us, by”, all of which are considered to be “stop” words [1].


Figure 3: Number of unique words

The words “children” and “childrens” both appear in the cloud. If stemming was performed, these two values would be stemmed down to “children” as shown in figure 4 via the nltk package in python. However, punctuations have been removed from the words as shown by the removal of the apostrophe from “children’s” to “childrens”.

```
>>> import nltk
>>> stemmer = nltk.PorterStemmer()
>>> print(stemmer.stem("children"))
children
>>> print(stemmer.stem("childrens"))
children
```

Figure 4: NLTK Porters Stemmer of children and children's

C.

The initial theory seems to be correct. The "stop words" are being removed and no "stemming" occurring was visual in another piece of text, as shown in figure 5. Words such as tweets, models and sentences are never stemmed showing two appearances with the single version of the word. The removal of words such as "you, but, also" has also been carried out.

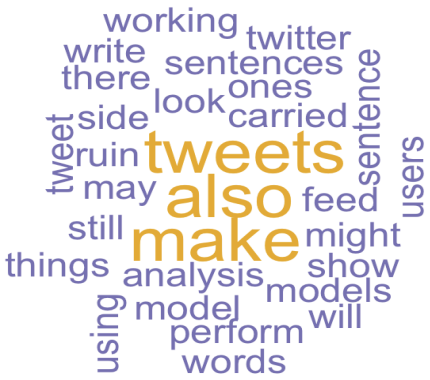


Figure 5: Word cloud on new text

D. When adding more words of "tweets" and "models" into the text, all other words disappear. This is evident in figure 8. This is now only showing the dominant words. Although this may be of benefit at times, for some tasks this may be considered a burden. For example, if a user was to scrape twitter using the "twitteR" package with the key words "earthquake + himalaya", then all the results would contain these two words. Using the "tm" package, these words can be removed after for a better result using the following code in figure 7.

```
> himEarth <- searchTwitter("earthquake+himalaya", n=50, resultType="recent")
```

Figure 6: Twitter scrapping "earthquake" and "himalaya"

```
> himClean <- tm_map(himClean, removeWords, c("himalaya","earthquake"))
```

Figure 7: Using TM package to remove the two words


Figure 8. Adding more repeated words.

Question 2.

A.

In the below graph, it can be seen that there are two peaks for the name "Mark Keane", one growing from 1969 to 1973 and another in 1993 to 2000. The latter is due to the increase in publications from Mark of which started to rapidly increase from 1993 as shown below in figure b. The first spike is due to the increase popularity of the first village manager of Oak Park in 1953 where "91% of voters were fed up with the Republican corruption" [3].


Figure 9A. Mark Keane Ngram

Costello, F. & Keane, M.T.; (1993) 'A model-based theory of conceptual combination' In: In K.Ryan & R. Sutcliffe (Eds.) Berlin: Springer-Verlag. (eds). *Artificial Intelligence and Cognitive Science '92.* , pp.7-15 [Details]

Keane, M.T. & Duff, S.; (1993) 'Towards an adequate cognitive model of analogical mapping' In: In H.Sorenson (Ed.) Amsterdam: Elsevier. (eds). *Artificial Intelligence and Cognitive Science.* [Details]

Andrews, J., & Keane, M.T.; (1993) 'A cognitive model of attention switching' In: In G. Orchard (Ed). Bristol & Philadelphia: IOP Publishing (eds). *Neural Computing Research and Applications.* , pp.139-144 [Details]

Costello, F. & Keane, M.T.; (1993) 'A model-based theory of conceptual combination' In: In K.Ryan & R. Sutcliffe (Eds.) Berlin: Springer-Verlag. (eds). *Artificial Intelligence and Cognitive Science '92.* , pp.7-15 [Details]

Smyth, Barry and Keane, Mark T (1994) 'Retrieving adaptable cases' In: Springer Berlin Heidelberg. , pp.209-220 [Details]

Smyth, Barry and Keane, Mark T (1995) 'Experiments on adaptation-guided retrieval in case-based design' In: Springer Berlin Heidelberg. , pp.313-324 [Details]

Figure 9B. Mark Keane Publications[2]

B.

Joe Kelly, shows one main peak in the 19th century between the years of 1823 to 1828. This is due to the famous musician, who wrote nine operas, four cantatas and numerous piano pieces. The other blip in the graph in the 1940s seems to be a combination of the Irish hurler and the formula 1 driver who ended his driving career because of the Oulton Park crash [4].


Figure 10. Joe Kelly Ngram

C.

In figure 11 below represents the introduction of the term “mobile telephone” into the English language. Before seeing this graph, I knew that commercial mobile phone was not created until the mid 1970s. Although this is correct and the first demonstration of a mobile phone didn’t take place until 1973 by Motorola, there are traces of the phrase being used long before it. Supposedly the race to create the first mobile phone began just after the second world war in the 1940s. This explains the large unexpected increase in the term years before the introduction of the technology. However, even before that in 1920s it can be seen to have a slight blip in the graph. In 1917 there was a patent filed for a “pocket-size folding telephone with a very thin carbon microphone” by a Finnish man named Eric Tigerstedt of which may contribute to this small slope [5].

Figure 11: Rapid growth of “mobile phone”

D.

The effects of “smoothing” the graph over different values can give misleading indications at times. This is because it will “spread” the result over the number of years indicated. However, often it is needed in order to get an understanding of the pattern being created. Below is a representation of the term “mobile telephone” using the smoothing value of 50, as can be seen this now indicates the term had been on a steady increase in use since the 1900s, which is obviously not the case as shown above in the figure of value 3. However, the “smoothing” effect is useful to show the trends, as the one above is given a value of 3. When comparing this to figure 12 b below, with a value of “0” and the result is a lot more volatile and difficult to visualize any pattern.

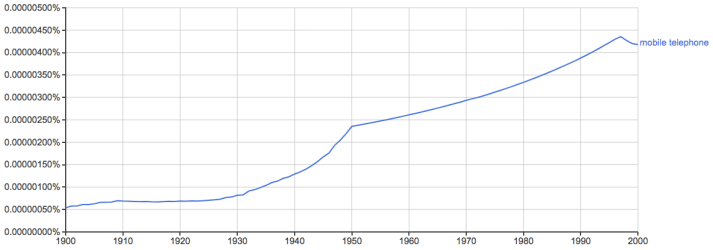

Figure 12 a: Smoothing value of 50

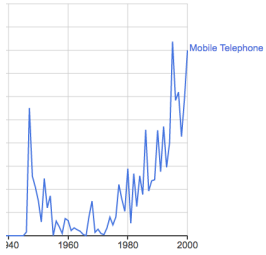

Figure 12 b: Smoothing value of 0

E.

An interesting comparison was to compare the generic term, “computer programming”, with different types of languages. Below it can be seen that the general term correlates significantly with the different languages as they make their mark in the industry. The main trend from 1980 to 1990 can be seen to increase with C programming and then decrease around the same year. When Java was introduced into the programming world in 1995, the general term begins to increase again with the language and then follow negative slope trend in the 2000s.

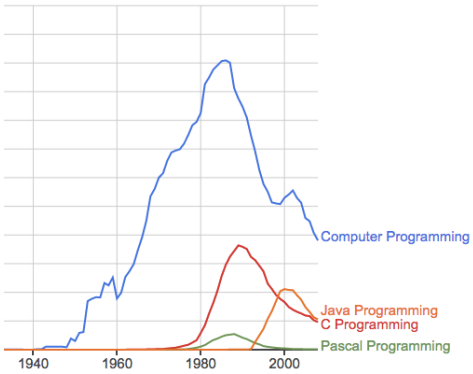

Figure 13: Comparing different words

F. Below is a graph of the word “record” being used as both a noun and a verb. The noun outperforms the verb significantly.

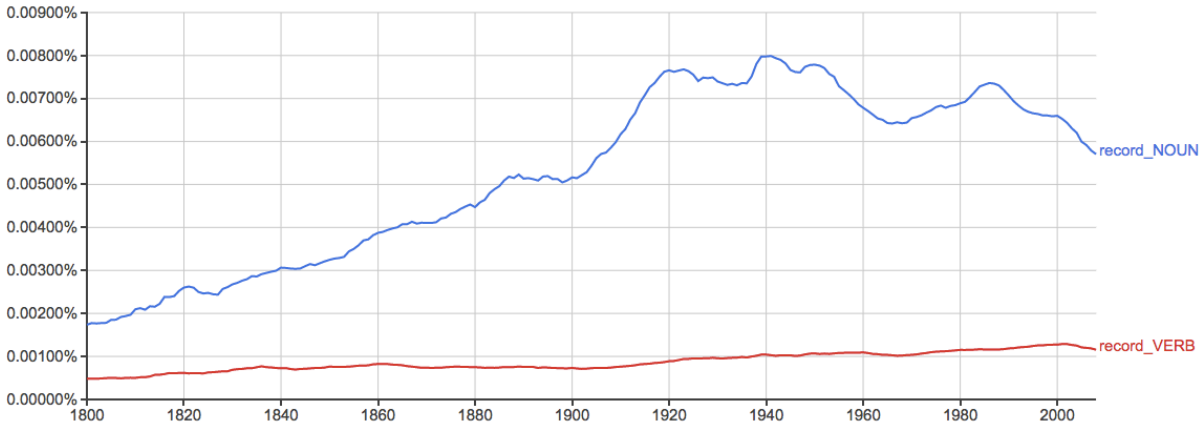


Figure 14: Record verb and noun

G.

An interesting topic today is around the word “guns”. The culture around the laws and regulations are constantly changing over time, especially in the past decade. When you search the term guns it can be seen that there are spikes within the years of the wars all the way back to the 1800s. This could be due to history documentations in which describe the types of guns used throughout the different periods of wars. What is interesting though, is how it has decreased significantly over the past number of years.


Figure 15: Guns N-gram

Although, when you compare this to the hot topic of “gun laws”, where over the past decade there has been significant pressure for governments in the US and other developed countries to regulate gun control better, it is inversely correlated with the word “gun” on its own. This term is also correlated significantly with the “2nd amendment” as expected shown in figure b.

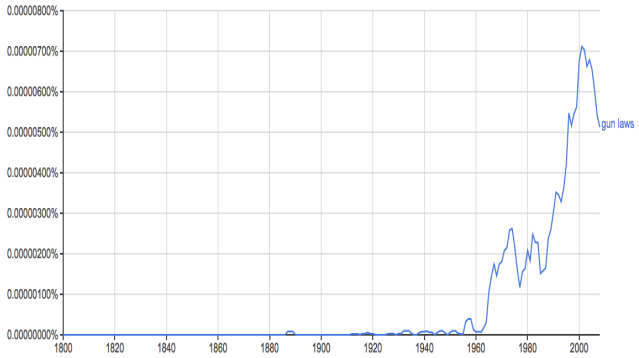


Figure 16 a: Gun Laws N-gram 2nd


Figure 16 b: amendment N-gram

## Question 3

Below is the Excel spread sheet of the normalized data. In the graphs below are the results of the different normalized data plots. This shows that no matter which method of normalization is chosen, even with different distribution of values, the trends will show the exact same pattern.

| | Raw Data | | | Normalised n | | | Normalised N | | | Difference | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Words | 2010 | 2011 | 2012 | 2010 | 2011 | 2012 | 2010 | 2011 | 2012 | 2010 | 2011 | 2012 |
| cat | 120 | 130 | 160 | 2% | 3% | 4% | 1% | 1% | 1% | 2% | 2% | 2% |
| dog | 200 | 220 | 260 | 4% | 5% | 6% | 1% | 2% | 2% | 3% | 3% | 4% |
| home | 150 | 140 | 100 | 3% | 3% | 2% | 1% | 1% | 1% | 2% | 2% | 2% |
| world | 120 | 175 | 195 | 2% | 4% | 4% | 1% | 1% | 1% | 2% | 3% | 3% |
| bird | 110 | 120 | 130 | 2% | 3% | 3% | 1% | 1% | 1% | 1% | 2% | 2% |
| apple | 200 | 150 | 100 | 4% | 3% | 2% | 1% | 1% | 1% | 3% | 2% | 2% |
| robber | 400 | 350 | 300 | 8% | 8% | 7% | 3% | 2% | 2% | 5% | 5% | 5% |
| king | 430 | 500 | 480 | 8% | 11% | 11% | 3% | 4% | 3% | 5% | 8% | 7% |
| queen | 600 | 640 | 550 | 12% | 14% | 12% | 4% | 5% | 4% | 8% | 10% | 8% |
| jim | 650 | 450 | 500 | 13% | 10% | 11% | 5% | 3% | 4% | 8% | 7% | 8% |
| hat | 800 | 500 | 650 | 16% | 11% | 15% | 6% | 4% | 5% | 10% | 8% | 10% |
| elephant | 950 | 700 | 600 | 19% | 16% | 13% | 7% | 5% | 4% | 12% | 11% | 9% |
| snake | 200 | 230 | 210 | 4% | 5% | 5% | 1% | 2% | 1% | 3% | 3% | 3% |
| orange | 120 | 120 | 150 | 2% | 3% | 3% | 1% | 1% | 1% | 2% | 2% | 2% |
| nail | 40 | 50 | 70 | 1% | 1% | 2% | 0% | 0% | 0% | 1% | 1% | 1% |
| Totals | 5090 | 4475 | 4455 | | | | | | | | | |
| N Total | 14020 | | | | | | | | | | | |

Figure 17: Excel calcualtions

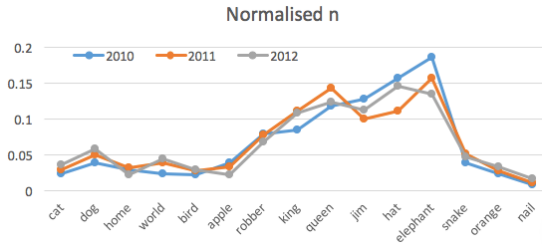


Figure 18 a: Normalized on year

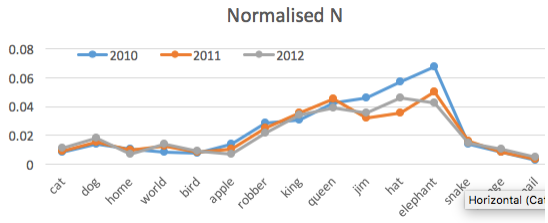


Figure 18 b: Normalized on all data

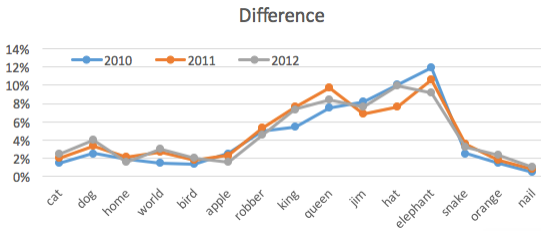


Figure 19: Normalized difference

## Question 4.

The paper indicates at the bottom of page (ii), that they use the package “CRAN.R”. This is a freely available open source statistic package. For this reason, once a CSV file from the google trends has been downloaded and real data of the same format used by Hyunyoung Choi and Hal Varian has been made available, this program would be possible to run. I have attempted for a number of hours to try and construct my own example. The format of which the script is reading in the ford data is unknown and has caused many errors in trying to get this to work. Even the format of the google trends data has changed since the code was constructed. My example was to try and compare the term “Iphone” in google trends with their quarterly sales shown below. Although unsuccessful, the results didn’t correlate with Hyunyoung Choi and Hal Varian conclusion. The reason for the google trends data unable to correlate with the sales was due to the company producing new phones every year. When Hyunyoung Choi and Hal Varian conducted their analysis on Ford, if the increase in the google searches was due to people looking at reviews to buy. While for the IPhone, the increase in searches spiked in September when new phones were being released but this didn’t necessarily mean people were interested in buying the new iphone, but interested in what was new. The largest sales were in Quarter

1 of each year which is January to the end of March, while the largest google trends were shown in September/October [7].

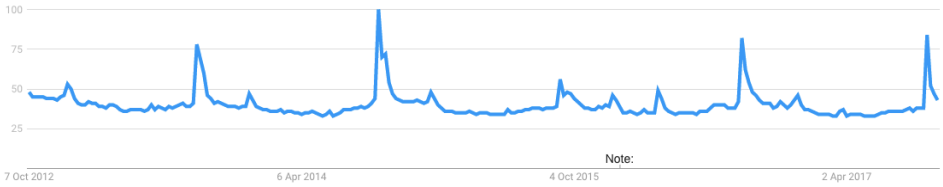

Figure 22: Google Trends IPhone

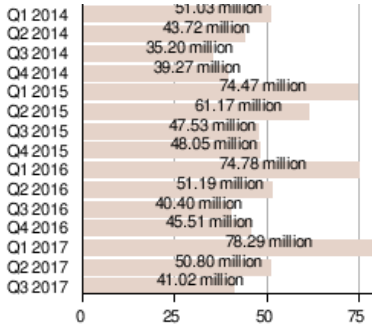

Figure 21: Iphone sales highest in first Quarter [6]

[1]. Accessed on 3/10/2017 https://nlp.stanford.edu/IR-book/html/htmledition/dropping-common-terms-stop-words-1.html
[2]. Accessed on 3/10/2017 http://www.ucd.ie/research/people/computerscience/professormarkkeane/
[3]. Accessed on 3/10/2017 http://www.oakpark.com/News/Articles/4-16-2013/Mark-Keane,-93,-Oak-Park's-first-village-manager/
[4] Accessed on 4/10/2017 https://gprejects.com/centrale/drivers/joe-kelly
[5]Accessed on 5/10/2017 https://www.motorola.com/us/about/motorola-history-milestones
[6] Accessed on 5/10/2017 https://upload.wikimedia.org/wikipedia/en/timeline/42dac140124701c6fc723dfdd3b147a0.png
[7] Hyunyoung Choi, Hal Varian "Predicting the Present with Google Trends", 2009.